DOCUMENT RESUME

ED 189 126 TM 800 292

AUTHOR Frary, Robert B.
TITLE Expected Multiple-Choice Test Item Scores Under Ordinal Response Modes.
PUB DATE Apr 80
NOTE 12p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Guessing (Tests): *Multiple Choice Tests: Responses: Scoring: *Scoring Formulas: Test Construction: Test Reliability: Test Validity
IDENTIFIERS Answer Until Correct: Distractors (Tests): *Ordinal Response Mode

ABSTRACT

Ordinal response modes for multiple choice tests are those under which the examinee marks one or more choices in an effort to identify the correct choice, or include it in a proper subset of the choices. Two ordinal response modes: answer-until-correct, and Coomb's elimination of choices which examinees identify as wrong, were analyzed for scoring rule appropriateness. Rules were analyzed in terms of expected item scores for different levels of information and misinformation, defined by the subset of choices believed to contain the answer. A general method for determining answer-until-correct scores was established. Answer-until-correct scores insured uniformity of guessing, which overcomes a major fault of the Coomb's mode and conventional correction-for-guessing scoring, namely, that the instructions are interpreted differently by examinees. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189126

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

EXPECTED MULTIPLE-CHOICE TEST ITEM SCORES UNDER ORDINAL RESPONSE MODES

by

Robert B. Frary

Abstract

Ordinal response modes are those under which the examinee selects one or more choices in an effort to identify the correct choice or include it in a proper subset of the choices. Answer-until-correct and the elimination response mode suggested by Coombs are examples of ordinal response modes and were selected for analysis in the present paper. Scoring rules for these response modes are analyzed in terms of expected item scores for different levels of information and misinformation defined according to the character of the subset of choices believed to contain the answer. A general method for determining answer-until-correct scoring rules is established.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. FRARY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800292

EXPECTED MULTIPLE-CHOICE TEST ITEM SCORES UNDER ORDINAL RESPONSE MODES

Robert B. Frary

Virginia Polytechnic Institute and State University

Ordinal response modes for multiple-choice tests are those under which the examinee marks one or more choices in an effort to identify the (single) right choice or at least include it in a proper subset of the choices. For example, on an answer-until-correct (AUC) test, examinees are instructed to continue selecting choices until the answer has been identified. Usually correctness of a choice is revealed when the examinee erases an answer sheet shield. This procedure gives immediate feedback to the examinee and is highly regarded for its potential for enhancing learning (Pressey, 1950). Another ordinal response mode is that established by Coombs (1953), under which examinees are instructed to mark those choices they firmly believe to be wrong. A score penalty for inadvertently marking the correct choice is invoked to promote compliance with the instructions. Of course, instructions to mark a single choice believed correct define a rudimentary ordinal response mode.

Development of the more complex ordinal response modes appears to have been motivated by two factors. The first is the belief or hypothesis that examinee test performance or learning will be enhanced by whatever cognitive processes are engendered by the response mode. The second, which is the main concern of this paper, is the hypothesis that score characteristics will be enhanced, especially reliability, through a more accurate determination of the examinee's knowledge on each item. While this hypothesis is plausible, empirical studies have yielded mixed results at best (see Hanna, 1977; Poizner, Nicewander and Gettys, 1978). In what follows, justification will be offered for the suggestion that this lack of success may be due in part to inappropriateness of the scoring rules associated with the response modes. In addition, a general method for establishing AUC scoring rules will be derived.

Existing Scoring Rules

For the AUC, Coombs and other ordinal response modes, it is possible to estimate from the examinee's response to an item the number of choices identified as wrong under two assumptions. These are that the examinee followed the instructions and was not misinformed, that is, the examinee did not classify the correct choice as a distractor. Item scores are then awarded based on the estimated number of distractors the examinee could identify. Under AUC, typically the examinee receives one point for each answer shield left unerased, which is equivalent to awarding credit for each item as a

linear function of the number of distractors identified (Brown, 1965). Exactly the same result follows for the usual method of scoring Coombs mode responses, under which the examinee receives one point for each correctly identified wrong choice. However, for an n-choice item, a penalty of n - 1 points is deducted for marking the right choice. These scoring rules yield expected item scores relatively much higher than does number-right scoring when the examinee can recognize some but not all distractors. At the same time, misinformation is penalized relatively more heavily when the examinee has narrowed the choices supposedly containing the answer to two or more distractors (see Frary, 1980).

Table 1 illustrates these circumstances for 4-item choices. As reported there, levels of information are defined as follows:

i_1 the correct answer is known

i_2 guessing is between two choices, including the answer

i_3 guessing is among three choices, including the answer

i_4 guessing is among all four choices.

Levels of misinformation are defined as follows:

m_1 only the correct choice has been categorized as wrong and guessing is among the three wrong choices

m_2 the right choice and one other have been categorized as wrong and guessing is among the remaining three wrong choices

m_3 the right choice and two others have been categorized as wrong, hence, the remaining wrong choice is believed correct.

TABLE 1 -- Expected Raw and Standardized[1] Item Scores Under the Information and Misinformation Levels for 4-Choice Items Using Number-Right and Conventional Coombs and AUC Scoring Rules

	Number Right		Coombs	AUC	Coombs/AUC
Level	Raw	Standardized	Raw	Raw	Standardized
i_1	1.00	1.00	3	3.0	1.00
i_2	.50	.33	2	2.5	.67
i_3	.33	.11	1	2.0	.33
i_4	.25	.00	0	1.5	.00
m_3	.00	-.33	-1	1.0	-.33
m_2	.00	-.33	-2	.5	-.67
m_1	.00	-.33	-3	0.0	-1.00

[1]Linear transformation to assign 1 to i_1, and 0 to i_4.

The values of Table 1 were calculated from simple probability relationships under the assumption of random guessing among subsets of choices containing the answer and, in the case of AUC levels of misinformation, the assumption that the examinee received no advantage from finding out that the answer was not in the subset initially believed to contain it and thereafter had to guess at random to find the answer.

It is not clear in every instance whether the original developers of the various ordinal response modes considered the desirability of rewarding partial information and penalizing misinformation differently from number-right scoring. It may be that the scoring rules adopted were simply ones that seemed fair and could be easily explained to examinees. There is certainly no reason to require scoring as described above, and there are examples of variations in the literature. Musser and Thompson (1977) scored 5-choice AUC items 10, 5, 2, 0, 0 for one through five erasures and offered no justification for this practice. Arnold and Arnold (1970) used the Coombs mode and assigned 5-choice item scores of 1, 3/8, 1/6, 1/16, 0 for successful marking of four through zero distractors, with a 1/4 point penalty for inadvertently marking the correct choice. This practice yields expected item scores linearly related to expected number-right item scores and was chosen for this purpose.

Establishing New Scoring Rules

Coincident with a decision to use an ordinal response mode must be the determination of the scoring rule. There may be substantial justification for giving credit on some basis other than as a linear function of the number of distractors identified. For example, it might be desirable to award very little partial credit for ability to recognize a single wrong choice, given the propensity of some item-writers to produce "throwaway" distractors. This outcome is essentially the case with number-right scoring, but not with AUC or Coombs as conventionally scored. In contrast, reliability and validity may be well served by the relatively high credit awarded by conventional Coombs or AUC scoring for ability to eliminate all but one wrong choice.

Given these considerations, it is important that expected item scores arising from any scoring rule be consistent with the intentions of the test constructor. With the Coombs mode, this result is easy to accomplish in the absence of misinformation and so long as examinees can and do follow instructions. If these conditions hold, the item score assigned for a response is also the expected item score. Then a penalty for marking the right answer must be determined so that guessing is penalized or at least neutralized. Of course, the penalty must also be such that the expected item scores for the levels of misinformation are acceptable given the estimated prevalence of misinformation and test content. For example, validity considerations might dictate high misinformation penalties when misinformation might be dangerous, as possibly in the case of a medical licensing examination. Another factor to consider in determining the penalty is whether it can, if necessary, be readily explained to examinees and whether they will consider it fair and reasonable.

For example, it might be judged desirable to award credit for correct identification of 0, 1, 2 or 3 wrong choices of a 4-choice item as follows under the Coombs mode:

0 wrong choice → 0
1 wrong choice → .1
2 wrong choices → .5
3 wrong choices → 1

Then, to neutralize gains from guessing, it would be necessary that potential losses for inadvertently marking the correct choice equal potential gains, or

P[marking right choice]·Penalty =
P[marking only wrong choice(s)]·Item score for no. of wrong choices marked. [1]

For guessing at random with one mark on a 4 choice item, the probability of marking the right choice is .25, and that of marking a wrong choice is .75. Then, with an item score of .1 for marking a single wrong choice (as proposed above), [1] yields a penalty of .3 (or an item, score of -.3) for marking only the right choice. Similarly, proposed item scores of .5 and 1 for correctly identifying two and three wrong choices yield item scores of -.5 and -.33 respectively for marking the right answer along with one and with two wrong choices.

It would be difficult to explain to examinees why the penalty for an error with two marks warrants a greater score deduction than for errors with one or three marks. Therefore, it might be better to adopt a uniform item score of -.5 when the correct choice is marked. Then it would be slightly less unattractive to guess at random with two marks than with one or three. This outcome may raise ethical questions in some quarters, but it need not yield adverse measurement effects if the penalty serves mainly to back up the admonition not to guess. Moreover, the higher penalty does not permit a guesser to gain an advantage (in the long run) over those who follow the rules. Finally, of course, under the Coombs mode, the assigned item scores are the expected item scores for the various levels of information and misinformation under the assumptions stated earlier.

Quite apart from the foregoing discussion, there are good reasons not to adopt the Coombs mode for a typical testing situation. Many investigators have found that examinees have great difficulty evaluating their levels of assurance as to whether a choice is right or wrong. Moreover, examinees may differ as to what constitutes a firm belief that a choice is wrong. Examinees often complain that the presence of the penalty for marking the right answer is emotionally taxing.

Expected Item Scores for Answer-Until-Correct Items

The deficiencies just mentioned clearly do not apply to AUC tests. However, the conventional AUC scoring rule may not be optimal from various

standpoints. Evidently some judgment to this effect motivated Musser and Thompson (1977) to adopt their 10, 5, 2, 0, 0 scoring rule for one through five erasures. However, simply adopting a revised scoring rule may not provide the expected item scores desired, since, unlike the case with Coombs mode item scores, there is not a very obvious correspondence between AUC assigned and expected item scores. Based on the same assumptions as Table 1, Table 2 shows expected raw and standardized item scores arising from Musser and Thompson's rule and compares them with number-right and conventional AUC expected item scores for 5-choice items.

TABLE 2 -- Expected Raw and Standardized[1] Item Scores Under the Information and Misinformation Levels for 5-Choice Items Using Number-Right, Conventional AUC and the Musser and Thompson (1977) AUC Scoring Rules.

	Number Right		Musser and Thompson AUC		Conventional AUC	
Level	Raw	Standardized	Raw	Standardized	Raw	Standardized
i_1	1.00	1.00	10.00	1.00	4	1.00
i_2	.50	.38	7.50	.62	3.5	.75
i_3	.33	.17	5.67	.34	3	.50
i_4	.25	.06	4.25	.12	2.5	.25
i_5	.20	.00	3.40	.00	2	.00
m_4	.00	-.25	1.75	-.25	1.5	-.25
m_3	.00	-.25	.67	-.41	1	-.50
m_2	.00	-.25	.00	-.52	.5	-.75
m_1	.00	-.25	.00	-.52	0	-1.00

[1] Linear transformation to assign 1 to i_1, and 0 to i_5.

One wonders whether Musser and Thompson worked out the results of Table 2 in deciding on their scoring rule. Regardless, it makes more sense to decide first on the desired expected values and then determine the scoring rule that corresponds. The procedure for doing so will be worked out for 4-choice items but can be applied for items with three or more choices.

Let x_1, x_2, x_3, x_4 be the item scores to be assigned for 1 through 4 erasures so that the expected item scores will be v_1, v_2, v_3, v_4. Let $v_1 = 1$ and $v_4 = 0$. Assign $x_1 = 1$ for a single erasure. Then the item score for two erasures is assigned on the basis that for each item with two erasures there was another for which v_2 was due but $v_1 = 1$ was received, or

$$v_2 = P[1 \text{ erasure}]x_1 + P[2 \text{ erasures}]x_2$$

$$= (.5)(1) + (.5)(x_2), \qquad [2]$$

from which

$$x_2 = 2v_2 - 1. \qquad [3]$$

Similarly, substituting [3] for x_2,

$$v_3 = P[1 \text{ erasure}]x_1 + P[2 \text{ erasures}]x_2 + P[3 \text{ erasures}]x_3$$

$$= (.33)(1) + (.33)(2v_2 - 1) + (.33)(x_3), \qquad [4]$$

or,

$$x_3 = 3v_3 - 2v_2. \qquad [5]$$

Finally, substituting [3] and [5] for x_2 and x_3,

$$v_4 = 0 = P[1 \text{ erasure}]x_1 + P[2 \text{ erasures}]x_2 + P[3 \text{ erasures}]x_3 + P[4 \text{ erasures}]x_4$$

$$= (.25)(1) + (.25)(2v_2 - 1) + (.25)(3v_3 - 2v_2) + (.25)(x_4), \qquad [6]$$

or,

$$x_4 = -3v_3. \qquad [7]$$

Note that under the restrictions adopted, negative raw item scores may be assigned and must be assigned for four erasures. For example, if $v_1 = 1$, $v_2 = .5$, $v_3 = .2$ and $v_4 = 0$ is desired, resulting x's are: $x_1 = 1$, $x_2 = 0$, $x_3 = -.4$, $x_4 = -.6$. A linear transformation of these x's may be desirable for presentation to examinees, such as: $x_1 = 8$, $x_2 = 3$, $x_3 = 1$, $x_4 = 0$.

While assignment of v_1, v_2, . . . is essentially arbitrary it is possible to choose these values so that $x_i > x_{i+1}$. Then, after identifying the answer with the ith erasure, the examinee would have to erase another shield to gain the maximum item score. Returning to the example just presented, suppose revised v's are adopted as follows: $v_1 = 1$, $v_2 = .5$, $v_3 = .1$, $v_4 = 0$. Resulting x's are: $x_1 = 1$, $x_2 = 0$, $x_3 = -.7$, $x_4 = -.3$. To avoid this problem, restrictions on the selection of v_2 and v_3 may be derived as follows:

$$x_2 < 1$$

$$x_3 < x_2$$

$$x_4 < x_3. \qquad [8]$$

Substituting [3], [5] and [7] in [8]:

$$2v_2 - 1 < 1 \text{ or } v_2 < 1$$

$$3v_3 - 2v_2 < 2v_2 - 1 \text{ or } v_3 < (4v_2 - 1)/3$$

$$-3v_3 < 3v_3 - 2v_2 \text{ or } v_3 < -v_2/3. \qquad [9]$$

The inequalities of [9] represent a simplex (Hadley, 1962), and their solution region is easily determined. Figure 1 shows the region of acceptable values for v_2 and v_3 which results. The point (.11, .33) represents the lowest values assignable to v_3 and v_2 respectively, and these are identical with the transformed expectation from 4-choice number-right scoring. Thus it is not surprising that $v_1 = 1$, $v_2 = .33$, $v_3 = .11$, $v_4 = 0$ applied in [3], [5] and [7] yield $x_1 = 1$, $x_2 = -.33$, $x_3 = -.33$, $x_4 = -.33$, which is a linear transformation of the usual 1 - 0 number-right scoring rule. Of course, the scoring rule $x_1 = 1$, $x_2 = -.33$, $x_3 = -.33$, $x_4 = -.33$ may be also viewed as related to correction-for-guessing scoring. However, at the item level, expected item scores under correction-for-guessing are a linear transformation of number-right expected item scores regardless of the extent to which examinees guess randomly (see Frary, 1980).

Figure 2 shows the three-dimensional region of acceptable values for v_2, v_3, v_4 for 5-choice items. Of course, simplexes for items with larger numbers of choices cannot be represented visually, though their solution regions are easily determined.

Discussion

At present, validity studies are lacking concerning the effect of allowing varying amounts of credit for partial information, and these need to be made across a wide variety of content areas. Also, studies need to be made concerning the high and unavoidable misinformation penalties resulting from conventional Coombs and AUC scoring. This feature may or may not be desirable from the standpoint of validity, perhaps depending on test-criterion content. Another lack is the technology for machine scoring of AUC responses, except in the case of tests administered interactively by computer. Measurement specialists with access to this medium should promote AUC testing as a first step toward familiarizing examinees and educators with this promising response mode. At the same time, optical mark reader manufacturers should begin product development work in this area.

With respect to computer administration of AUC tests two interesting observations may be made:

1) It would be possible to incorporate into the scoring process weighting factors based on the sequence and difficulty level of distractors selected prior to the answer.

2) The problem of unacceptable expected item scores (outside the solution region of the related simplex) could be avoided since, the computer need not accept additional responses after identification of the answer. Of course, it would then be almost a requirement that examinees not be told the scoring rule.

Although substantial development work needs to be done with respect to AUC, its reported usage by Hanna and others has yielded some positive results. Compared with number-right scoring, higher reliability estimates have been reported, and student acceptance has been good. Most important, AUC tests virtually insure uniformity of guessing behavior, which overcomes a major fault of using the Coombs mode and the conventional correction-for-guessing, namely, that the instructions are interpreted in various ways by examinees (see Cross and Frary, 1977). Thus, there is reason to believe that AUC scoring offers substantial promise for improving measurement quality.

REFERENCES

Arnold, J. C., & Arnold, P. L. On scoring multiple choice exams allowing for partial knowledge. The Journal of Experimental Education, 1970, 39(1), 8-13.

Brown, J. Multiple response evaluation of discrimination. The British Journal of Mathematical and Statistical Psychology, 1965, 18(1), 125-137.

Coombs, C. H. On the use of objective examinations. Educational and Psychological Measurement, 1953, 13, 308-310.

Cross, L. H., & Frary, R. B. An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Frary, R. B. The effect of misinformation, partial information and guessing on expected multiple-choice test item scores. Applied Psychological Measurement, in press.

Hadley, G. Linear Programming. Reading, Mass.: Addison-Wesley, 1962.

Hanna, G. S. A study of reliability and vlaidity effects of total and partial immediate feedback in multiple-choice testing. Journal of Educational Measurement, 1977, 14, 1-7.

Musser, G., & Thompson, L. A learning center based system of instruction. The American Mathematical Monthly, 1978, 85, 290-293.

Poizner, S. B., Nicewander, W. A., & Gettys, C. F. Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. Applied Psychological Measurement, 1978, 2(1), 83-96.

Pressey, S. L. Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. The Journal of Psychology, 1950, 29, 419-447.

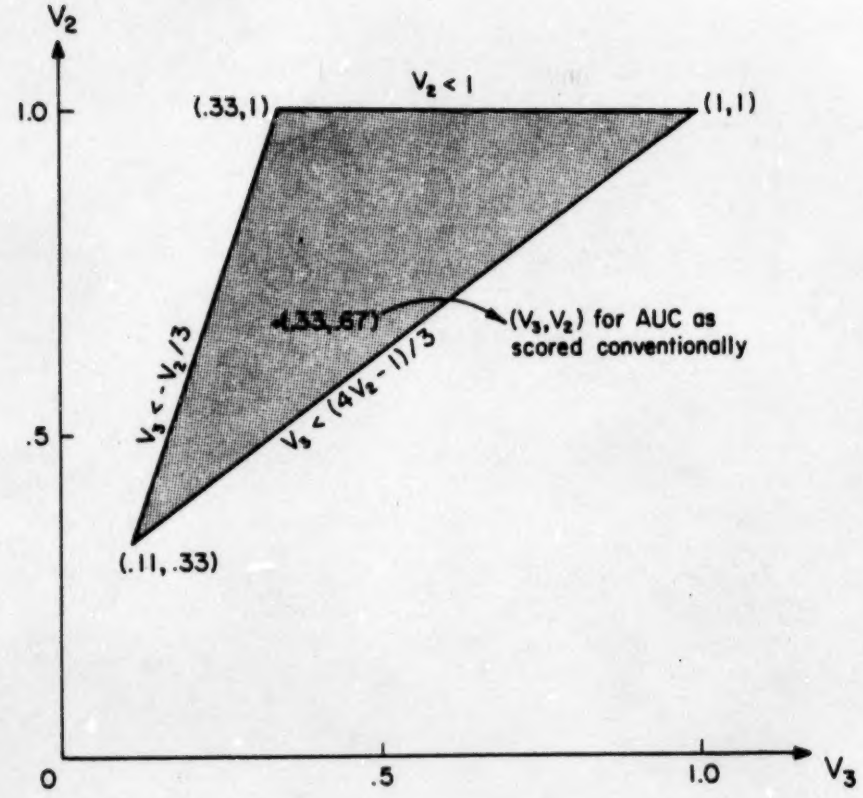

Figure 1 - Region of Acceptable Expected Values (v_3, v_2) for Establishing AUC Scoring Rules for 4-Choice Items .

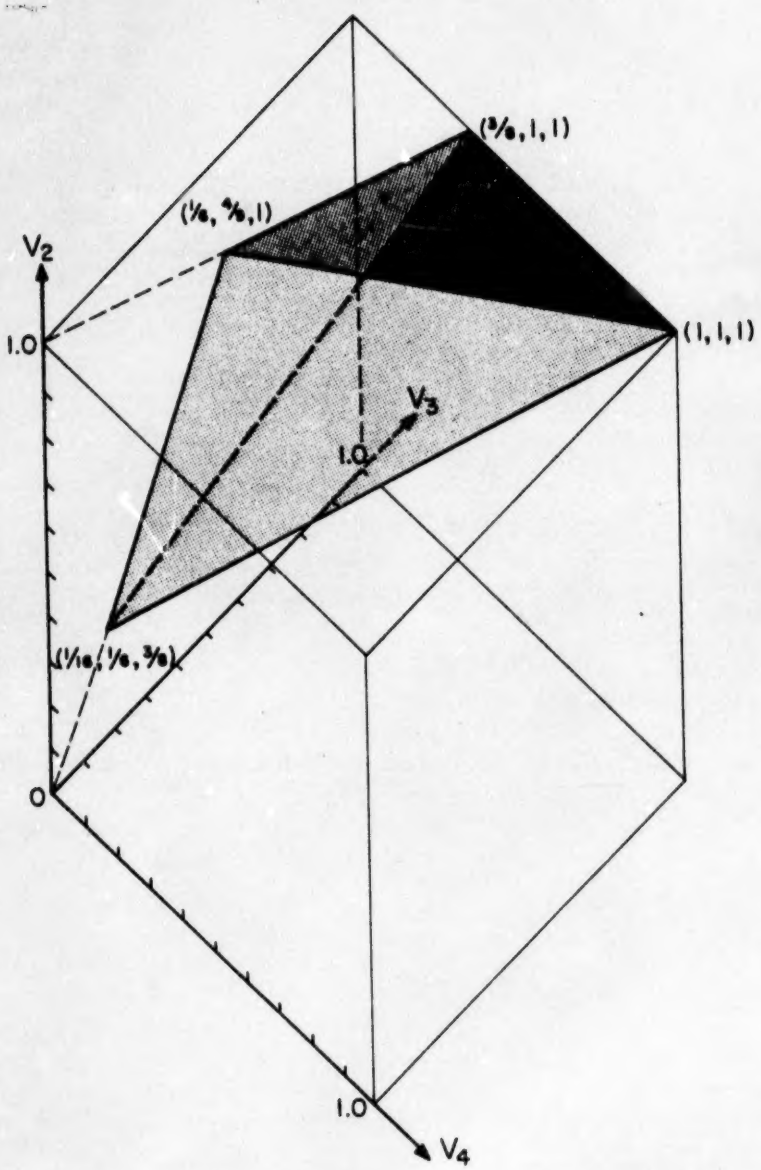

Figure 2 - Region of Acceptable Expected Values (v_4, v_3, v_2) for Establishing AUC Scoring Rules for 5-Choice Items